Speech-To-Text Evaluation Task Proposal

1. Champion: Jon Fiscus, jfiscus@nist.gov

2. Task name: Speech-To-Text Transcription

3. Task description: Transcribe all the words spoken in a meeting excerpt.

4. Task data type attributes and allowable attribute values:
   Systems output Conversation Time Mark (CTM) records which indicate the meeting ID/source file, audio channel, begin time, duration, lexical orthography, confidence score, and lexical type. See the "Appendix B: Conversation Time Mark (CTM) Format STT System Output" of the "RT-05S Evaluation plan" for details.

5. Justification: Speech-To-Text transcription is a major source of information to begin understanding the content and structure of a meeting. The meeting recognition community has worked on this task for several years and this proposal seeks to continue that work.

6. Rules for annotation:
   The evaluation test set will be transcribed according to the MeetingDataCarefulSpec-V1.2 spec. at the URL http://www.ldc.upenn.edu/Projects/Transcription/NISTMeet/MeetingDataCarefulSpec-V1.2.pdf.

7. Required core technologies: None

8. Suggested metric: Word Error Rate (WER) will be the primary metric. WER is an overall STT error score computed as the average number of token recognition errors per reference token:

   $$Error_{STT} = \left(N_{Del} + N_{Ins} + N_{Subst}\right) / N_{Ref}$$

   Where

   N_{Del} = the number of unmapped reference tokens,
   N_{Ins} = the number of unmapped STT output tokens,
   N_{Subst} = the number of mapped STT output tokens with non-matching reference spelling per the token rules above, and
   N_{Ref} = the maximum number of reference tokens

   The rules defining system output and scoreable tokens will be identical to the rules specified in the RT-05S evaluation plan. The only exception will be the scoring of simultaneous overlapping speech. Systems running with a distant microphone audio condition will be scored on non-overlapping speech and simultaneous speech with up to 5 active participants per segment of the meeting. The new SCTK alignment module will be used to score STT system output.

9. Suggested evaluation conditions: The audio input conditions for the STT will change from RT-05S to RT-06S. The audio input conditions will be:

    a. All Distant Microphones (ADM): All microphones not worn by the participant will be used for this condition. This condition includes microphones on the periphery of the room and the centrally located microphones.
    b. Central Distant Microphones (CDM): All microphones placed in between the participants in some central location. The determination of "central" is made during the opening of the meeting and therefore does not change during the course of the meeting as participants move. This audio condition was formerly called Multiple Distant Microphones (MDM).
    c. Individual Head Microphones (IHM): Each participant located in the meeting room must be equipped with a high-quality microphone head set and therefore a separate channel will be provided for each meeting participant. Meeting participants not located in the room, e.g. teleconference participants, will not be considered transcribable speech for this condition and therefore not be scored.

    The primary condition will no longer be the CDM (formerly the MDM condition). Instead, the primary condition will be the ADM condition.

10. Applicable domain:Any domain supported by the STT task.

11. Required resources: Development test corpora are already available

12. Other notes and references to existing or contributing work:
    See the RT-05S evaluation plan
    http://www.nist.gov/speech/tests/rt/rt2005/spring/rt05s-meeting-eval-plan-V1.pdf